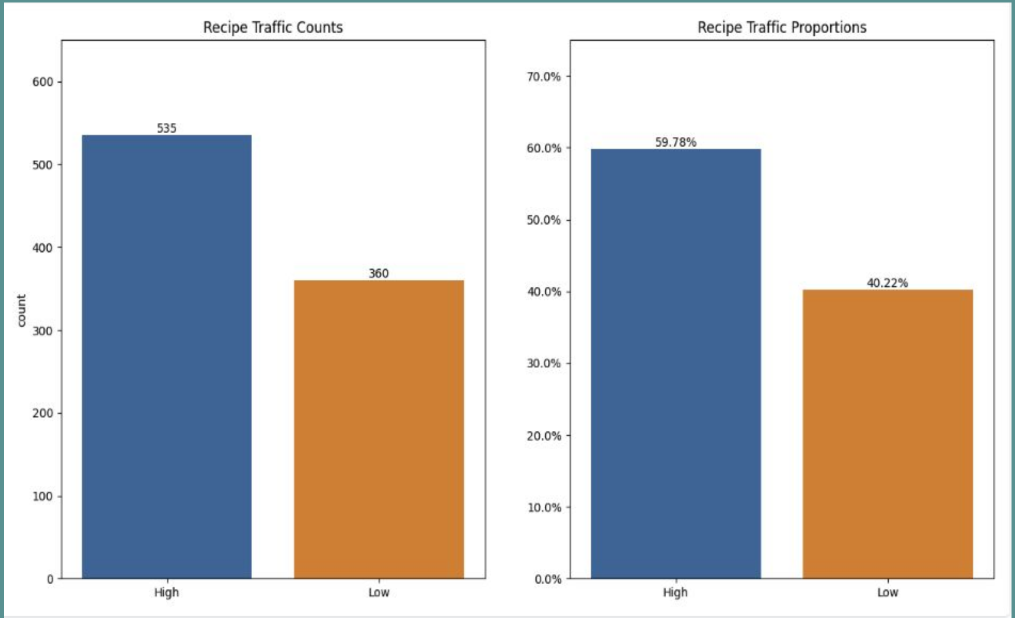
Tasty Bytes Recipe Traffic Classifier

Customer Subscription Generation

Business Goals & Analysis Objectives

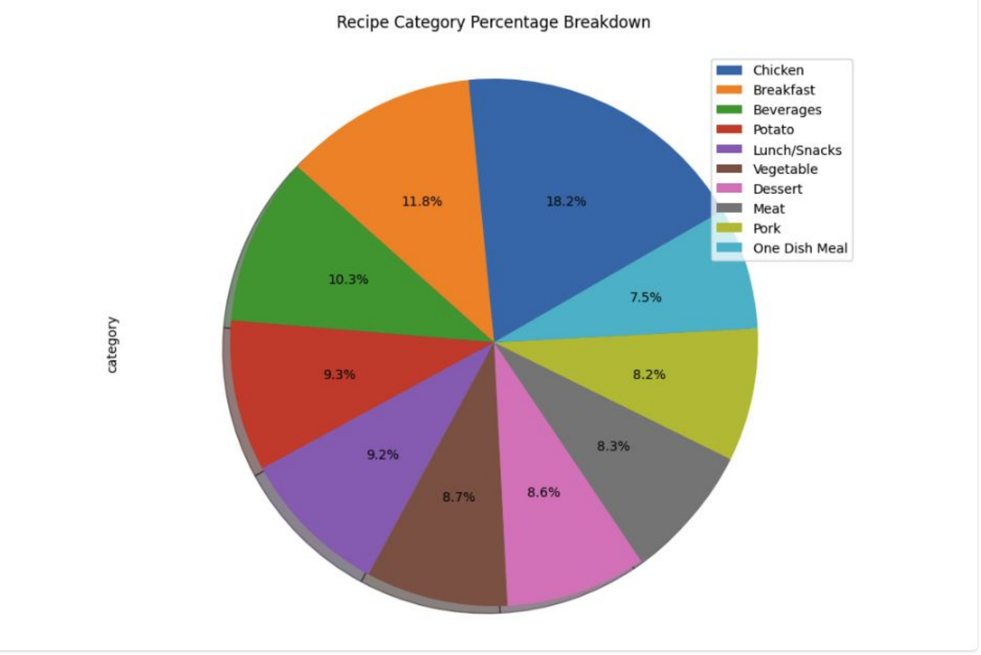
- Business Goals & Metric
  - Identify & predict recipes that will generate higher website traffic
  - Use increased web traffic to generate further customer subscriptions
  - Correctly predict high traffic recipes 80% of the time
  - **Metric** - F1 Score From Models Predictions
- Analysis Objectives
  - Detail & visualize relationships between recipe characteristics and their correlation to greater traffic
  - Fit, tune multiple classification models to predict a recipe's traffic generation
  - Review selected classification models predictive power and overall classification report for insight on **precision**, **recall** and overall **accuracy** for traffic labels (high & low)

Target Traffic : Recipe Counts/Proportions



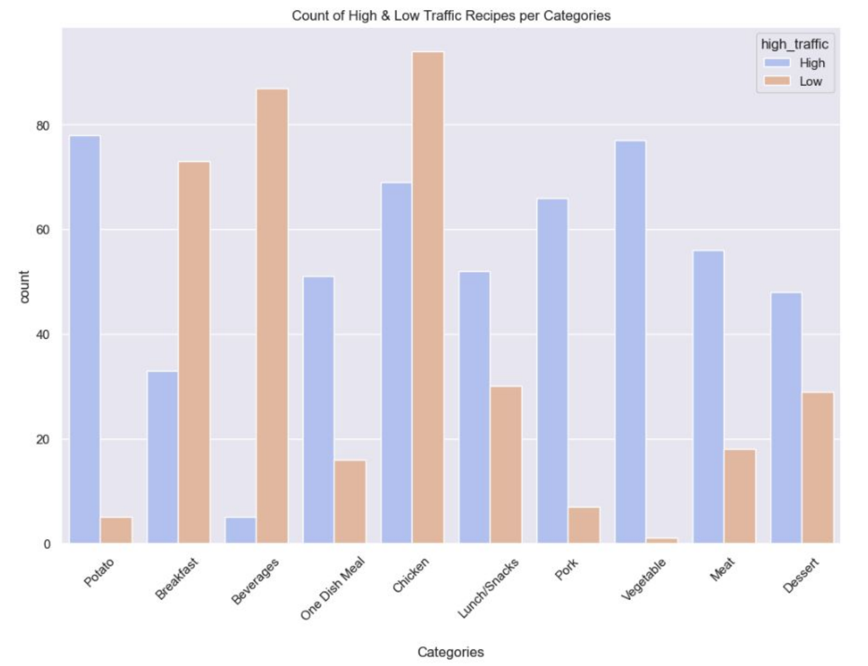
- **After cleaning and preparing the data for analysis, the provided traffic counts showed us that of the available recipes shared, nearly 60% resulted in high_traffic.**

Recipe Category Makeup



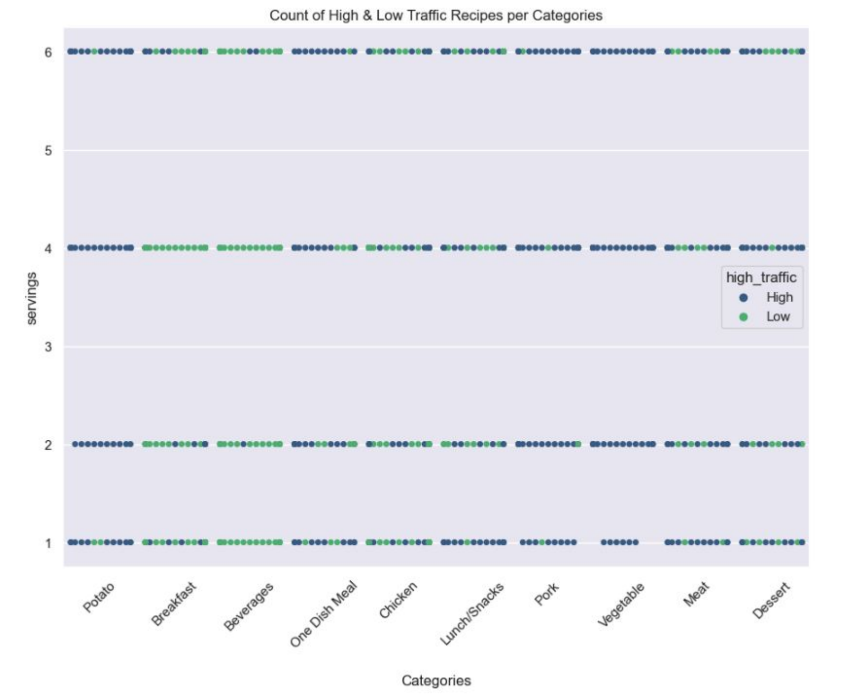
- To better understand what type of recipes currently populate the homepage, we can use the following visual to establish a baseline for how often a recipe from the following categories appears

Outcomes : Traffic By Category



- As our classification model will look to predict recipes' popularity and ability to drive traffic, we can use the following graphic to get an indication on how they categorize currently
- We can see a few categories initially (Beverages & Breakfast) which very rarely result in high_traffic whereas recipes in (Potato, Pork, Vegetables, & Meat) that have quite good traction in generating high_traffic
- This does beg the question, how does Serving Size impact the outcome if at all

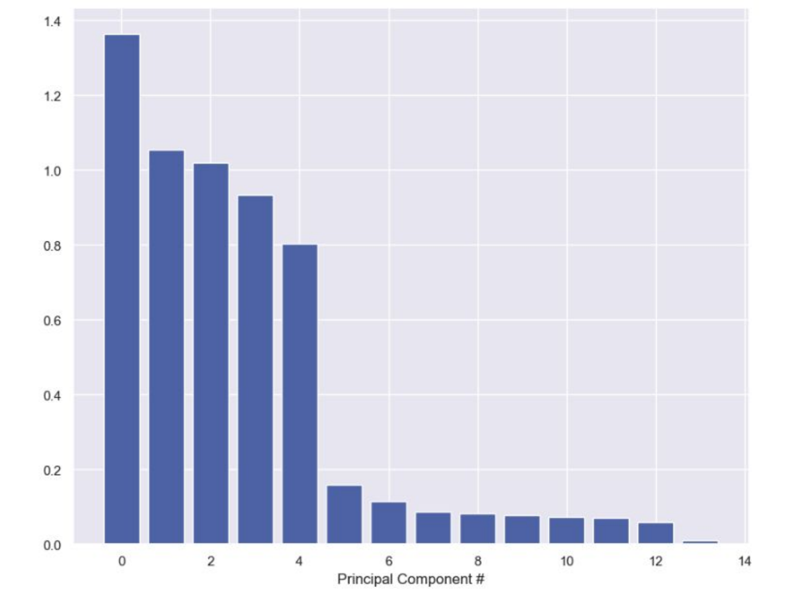
Outcomes : Traffic By Category & Serving Size



- In conjunction with the previous slide, we can see from the traffic legend how difference recipe categories scores across serving sizes in either attracting "high" or "low" traffic
- With the counts showing the overall trend for the categories this allows us to look at categories with more balanced representation for recipes drawing high or low traffic
    - Ex - Lunch/Snacks
        - While having a positive ratio for more high traffic recipes, we see distinct proportion differences across the serving sizes

Classification Model(s) : PCA

- **Principal Component Analysis - Feature Results**
  - Although 5 features represented a good "elbow" point in explaining Variance for our target variable of "Traffic" the last model fit and recommended to the business ended using 13 features (binary "food" categories encoded for Model inclusion)
  - The "continuous" values for the model's development (recipe nutritional value - ex. Carbs) were most definitive in explaining variance
  - Recipe's category features alone have small variance but cumulatively summed did help improve model accuracy by ~6-8% when included



LogReg & SVM - Model Parameters & Accuracy

- Shared Model Fitting Processing & Validation
  - With the assistance of GridSearch for hyperparameter tuning and KFold model train/test splitting we were able to see which of the classifier estimators was more accurate in predicting traffic by a recipe
- Logistic Regression - Results
  - PCA - 5
    - Tuned Logistic Regression Parameters: {'log_reg__C': 0.2113157894736842, 'log_reg__fit_intercept': True, 'log_reg__penalty': 'l2', 'log_reg__solver': 'newton-cg'}, Accuracy: 0.6875
  - PCA - 13
    - Tuned Logistic Regression Parameters: {'log_reg__C': 0.05357894736842105, 'log_reg__fit_intercept': False, 'log_reg__penalty': 'l2', 'log_reg__solver': 'newton-cg'}, Accuracy: 0.7410714285714286
- SVM - Results
  - PCA - 5
    - Tuned Logistic Regression Parameters: {'svm__C': 0.47421052631578947, 'svm__decision_function_shape': 'ovo', 'svm__gamma': 'auto', 'svm__probability': True}, Accuracy: 0.6741071428571429
  - PCA - 13
    - Tuned SVC Parameters: {'svm__C': 0.9474210526315789, 'svm__decision_function_shape': 'ovo', 'svm__gamma': 'scale', 'svm__probability': True}, Accuracy: 0.75

Model Selection - Classification Report

- <u>Classification Report</u>
  - With the defined <u>F1</u> score from as the metric for model selection, the slight advantage with the SVM model on the "weighted" accuracy of the metric would lead us to choose this model.

Log Reg Report

	precision	recall	f1-score	support
0	0.59	0.70	0.64	73
1	0.84	0.76	0.80	151
accuracy			0.74	224
macro avg	0.71	0.73	0.72	224
weighted avg	0.76	0.74	0.75	224

SVM Report

	precision	recall	f1-score	support
0	0.62	0.62	0.62	73
1	0.81	0.81	0.81	151
accuracy			0.75	224
macro avg	0.72	0.72	0.72	224
weighted avg	0.75	0.75	0.75	224

Business Recommendations

- **Model Testing vs Current Product Recipe Selection**
  - A/B Testing for the current recipe selection and the deployed model recipe selection could be reviewed for 6-10 weeks of data collection to see if the model's selection lead to higher traffic and ultimately higher customer subscriptions
- **Model Improvement & Further Data Collection**
  - Data Collection & Completeness
    - The PCA analysis detailed the importance of the nutritional values for model explained variance.
    - 52 recipes did not include any of this data which resulted in recipe exclusion from the model
  - Model Improvement
    - Our classification report detailing the "precision" and "recall" for **low_traffic** recipe classification was significantly lower than **high_traffic** recipes
    - Further recipes resulting in **low_traffic** should aid in the model's predictive overall metrics for each traffic classification type